Biomarkers for Mixed Dementia: a hard bone to bite? Preliminary analyses and promising results for a debated topic*

Andrea Campagner[1,*,†], Lorenzo Famiglini[2,†], Beatrice Arosio[3], Paolo Rossi[4], Giorgio Annoni[5] and Federico Cabitza[2,1]

[1] *IRCCS Istituto Ortopedico Galeazzi, Milan, Italy*

[2] *University of Milano-Bicocca, Milan, Italy*

[3] *Department of Clinical Sciences and Community Health, University of Milan, Via della Commenda 19, 20122 Milan, Italy*

[4] *General Medicine, Hospital San Leopoldo Mandic, Largo Mandic, 1, 23807, Merate (Lecco), Italy*

[5] *Dipartimento di Medicina , Università di Milano-Bicocca, Milan, Italy*

Abstract

Dementia refers to a group of neurodegenerative disorders that impact the cognitive function of an increasing number of individuals. Because of the variety of manifestations, the idea of *mixed dementia* has recently garnered increased awareness and attention from the scientific community. In this work, we describe a high-quality dataset, as well as the findings of a preliminary analysis devoted to investigating the potential of computational methods that are highly indicative of mixed dementia. We will specifically describe the findings of a phenotypic stratification analysis, based on clustering approaches, that highlights possibly significant aspects of mixed dementia, paving the way for further research devoted to the application of Machine Learning techniques to the robust and early diagnosis of mixed dementia.

Keywords

mixed dementia, dementia, medical Artificial Intelligence, clustering analysis

1. Introduction

Dementia is a term that encompasses a range of neurodegenerative diseases that affect the cognitive function of a growing number of patients worldwide, in the order of tens of millions, as the global population ages. Among the diverse subtypes of dementia, Alzheimer's disease (AD) and vascular dementia (VaD) are the most prevalent [1, 2] but Lewy body dementia can occur either alone or in any combination with the above mentioned types of conditions. In light of this heterogeneity of manifestations, recently the concept of *mixed dementia* has received

AIxAS 2023: Fourth Italian Workshop on Artificial Intelligence for an Ageing Society, 6–9 November 2023, Rome, Italy

[*]

[*] Corresponding author.

[†] These authors contributed equally.

andrea.campagner@unimib.it (A. Campagner); l.famiglini@campus.unimib.it (L. Famiglini); beatrice.arosio@unimi.it (B. Arosio); paolodionigi.rossi@gmail.com (P. Rossi); giorgio.annoni@unimib.it (G. Annoni); federico.cabitza@unimib.it (F. Cabitza)

increasing recognition and attention by the scientific community: mixed dementia is a condition characterized by the *coexistence* of features of more than one form of dementia (typically AD and VaD). Clearly, due to the overlap of symptoms among its constituent conditions, diagnosing mixed dementia poses a unique set of challenges. Moreover, conventional diagnostic tests often cannot distinguish mixed dementia from its individual components, and this can lead to misdiagnosis and inappropriate treatment plans [3]. The complex pathology of this condition further complicates the interpretation of biomarkers and neuroimaging studies, which may reflect aspects of multiple underlying diseases [4]. Moreover, recent studies [5] have cast uncertainty upon some established points in the scientific community, regarding the role of plaques in brain tissue as a primary cause of the illness (the so-called amyloid hypothesis) and the importance of imaging in the diagnosis of dementia.

For all these reasons, finding new and better ways to early and accurately diagnose mixed dementia is pivotal for effective treatment and care. For instance, traditional interventions for AD, such as cholinesterase inhibitors, have demonstrated limited efficacy in treating mixed dementia, underlining the need for targeted treatments [6]. Moreover, an early diagnosis enables timely management of vascular risk factors, which could attenuate the progression of cognitive decline. In light of these challenges, there is a strong need for the development of reliable diagnostic methods. Recent advancements in the fields of machine learning and computational neuroscience offer promising avenues for creating robust algorithms capable of classifying complex cases of mixed dementia by potentially integrating a wide range of data, including neuroimaging, genomics, and clinical assessments.

This manuscript aims to present how a unique and high-quality dataset, collected at the Policlinico Hospital of Milan (Italy), can be exploited to explore the potential of computational methods such as predictive computing or clustering techniques, in detecting biomarker patterns that are unique to, or highly indicative of, mixed dementia. In particular, we will present the results of a preliminary analysis, based on the above-mentioned dataset, aimed at the description and phenotypic stratification of mixed dementia. By using clustering techniques and statistical inference methods for doing so, we aim to contribute to filling a critical gap in the literature by focusing on the early detection, classification, and phenotypic stratification of mixed dementia, thereby facilitating targeted early detection and personalized therapeutic interventions for this complex and highly impactful disease of our times.

2. Methods

In this section, we describe the methodology adopted for the descriptive analysis and phenotypic stratification of the dataset. The dataset for this study was collected at the Policlinico Hospital of Milano (Milan, Italy) and encompassed 911 records (for as many single patients), collected between March 2009 and March 2018. The dataset encompassed a total of 71 columns, including demographics, co-morbidity, clinical information as well as laboratory parameters. The full list of features is reported in Tables 1 and 2. In particular, in the dataset, each patient was associated with one among seven diagnoses of dementia (not including the control group), namely: Alzheimer's Disease (AD), Parkinson's Disease (PD), Dementia with Lewy's Bodies (LB), Frontotemporal Dementia (FTD), Mild Cognitive Impairment (MCI), idiopathic Normal

Pressure Hydrocephalus (iNPH), Mixed dementia (MD).

2.1. Preprocessing

The initial step in preprocessing involved removing missing or duplicate entries from the dataset, which initially comprised 911 records. Columns that were considered not relevant for clinical purposes, such as practice number and date of entry, were excluded from further analysis.

Missing data were addressed by eliminating rows and subsequently, columns with at least 40% missing values, resulting in a dataset of 906 instances and 61 columns. Given the dataset's combination of discrete and continuous variables, we employed different strategies for handling remaining null values. For discrete variables, instances with missing values were removed, while continuous variables were imputed using the MICE (Multivariate Imputation by Chained Equations) algorithm [7]. This yielded a dataset of 899 instances and 61 variables.

Outlier detection and removal were performed using the Isolation Forest algorithm [8], leading to the identification and elimination of 90 outliers.

2.2. Dimensionality reduction and Clustering

Since the dataset encompassed both discrete and continuous features, the Gower similarity coefficient [9] was employed to calculate distances among instances in the dataset. Dimensionality was subsequently reduced using the t-SNE algorithm [10], setting the perplexity hyper-parameter to 20, and the number of iterations to 3500. We decided to use t-SNE, rather than other dimensionality reduction methods (e.g., PCA), as it allows to flexibly model non-linear relationships among features.

HDBSCAN clustering [11] was applied to the reduced data: specifically, HDBSCAN was applied to the output of the t-SNE dimensionality reduction step. We decided to use HDBSCAN as, being a density-based algorithm, it does not require the specification of a fixed number of clusters, but rather allows the automatic discovery of the number of groups in the data. Specifically, we decided to use HDBSCAN rather than DBSCAN or OPTICS as it does not require the specification of a distance threshold, which is automatically inferred through the application of a hierarchical clustering algorithm, while being less prone to the identification of noisy clusters. Cluster quality was assessed by visual inspection and analysis of the features' distributions in the different clusters (see next Section). Hyper-parameters of both t-SNE nad HDBSCAN (in particular, perplexity) were selected so as to optimize the Silhouette score of the resulting clustering (thus, the criterion for hyper-parameter selection was based on a purely internal criterion, with no reference to the dementia-type labeling of patients).

2.3. Statistical Testing and Correction

To examine the characteristics of patients in different clusters, and understand if these correlated with different forms of dementia, we applied various statistical tests to compute p-values for both discrete and continuous variables within these clusters.

For discrete variables, we used the proportions Z-test. By contrast, for continuous variables, the non-parametric Mann-Whitney U test was employed. Considering that multiple tests were performed across different variables, we applied the Benjamini-Hochberg false discovery rate correction method to adjust the p-values, which allows to control the false discovery rate among a family of related hypothesis tests.

Feature	Distribution	Missing
Sex	F: 64%, M: 36%	
Marital Status	Married: 54%, Widowed: 36%, Unmarried: 7%, Divorced: 3%	<1%,
Education	Elementary school: 33%, High school: 26%, Middle school: 26%, University: 12%, Illiterate: 1%	1%
Living alone?	No: 65%, Yes: 35%	
Caregiver?	No: 84%, Yes: 16%	
Smoking	No: 55%, Stopped: 35%, Yes: 9%,	2%
Hypertension?	Yes: 67%, No: 33%	
Diabetes	No: 85%, Yes: 15%	
Heart failure	No: 96%, Yes: 4%	
Ischaemic Heart Disease	No: 85%, Yes: 15%	
Arrhythmia	No: 72%, Yes: 28%	
COPD	No: 87%, Yes: 13%	
Hypovisus	1: 53%, 0: 47%	
Hearing loss	No: 73%, Yes: 27%	
Arthrosis	No: 54%, Yes: 46%	
Atherosclerosis	No: 51%, Yes: 49%	
Renal failure	No: 90%, Yes: 10%	
Liver disease	No: 83%, Yes: 17%	
Anxiety/Depression	No: 53%, Yes: 47%	
Cerebrovascular disease	Yes: 60%, No: 40%	
Cognitive impairment Malignant tumor	No: 55%, Yes: 45%, No: 81%, Yes: 19%	
Osteoporosis	No: 79%, Yes: 21%	
Metabolic syndrome	No: 95%, Yes: 5%	
Anemia	No: 91%, Yes: 9%	
Diverticula	No: 83%, Yes: 17%	
Pain	No: 72%, Yes: 28%	
Alterations in bowel habits	No: 68%, Yes: 32%	<1%
Sleeping disorders	No: 68%, Yes: 32%	< 1%
Heart murmurs	Yes: 52%, No: 48%	
Abnormal gait	No: 64%, Yes: 36%	
Peripheral edema	No: 73%, Yes: 27%	< 1%
Trembling	No: 92%, Yes: 8%	<1%,
Dementia	No: 27%, MCI: 31%, MD: 26 %, AD: 9%, iNPH: 4%, FTD: 1%, LB: 1%, PD: <1%	

Table 1
Distributions of the categorical features

3. Results

The distributions of both the discrete and continuous features are summarized in Tables 1 and 2. Mixed dementia occurred in approximately 1 out of 4 patients. The dataset was strongly

Feature	Mean	St.Dev.	Missing
Age	79.6	5.9	0%
BMI	25.7	4.5	23%
Saturation (%)	96.9	5.3	18%
MMSE	24.1	5.3	3%
FI	0.3	0.1	1%
Hemoglobin	13.4	1.4	12%
Red blood cells	4.5	0.5	18%
White blood cells	23.7	350.7	16%
Platelets	229.4	65.5	15%
MCV	89.9	7.1	15%
MCH	30.0	2.4	18%
Creatinine	0.9	0.3	14%
Glucose	98.2	27.4	17%
Na	141.8	2.8	20%
K	4.4	0.4	19%
Transferrin	253.1	44.7	56%
Ferritin	145.1	129.4	49%
Vitamin B12	415.8	489.5	21%
Folates	7.7	7.7	21%
TSH	2.5	2.6	22%
PCR	0.6	2.0	44%
Treponeme	0.4	1.5	71%
Systolic blood pressure	141.8	17.9	0%
Diastolic blood pressure	79.1	9.8	0%
Clock Drawing Test	2.8	2.0	8%
GDS	14.2	48.0	53%
Tinetti POMA	20.3	6.0	43%
AST GOT	21.1	9.5	23%
ALT GPT	17.5	10.1	22%
Blood iron	83.5	27.9	49%
Cholesterol	205.3	39.4	24%
Vitamin D	18.3	14.2	30%
ADL	4.9	1.3	6%
IADL	4.9	2.5	6%

Table 2
Mean, standard deviation, and missing rate for the continuous feature.

imbalanced in terms of both sex (females had about twice the frequency of males): this imbalance, however, is consistent with the literature on dementia [12]. We also found a skewed age distribution (average age higher than 75), which is consistent with dementia being an aging-related spectrum of diseases. The results of the outlier analysis are reported in Figure 1.

The results of the t-SNE dimensionality reduction step are reported in Figure 2: patients who were associated with mixed dementia clustered separately from other patients.

The results of the clustering analysis are reported in Figure 3. HDBSCAN identified three different clusters (as well as regions of noise points surrounding and separating them): in particular, cluster 1 (teal color in Figure 3) was strongly associated with patients diagnosed with mixed dementia. The coverage of the clusters was 75% (approximately 1 instance out of 4 was classified as a noise point), while the DBVC score was 0.4.

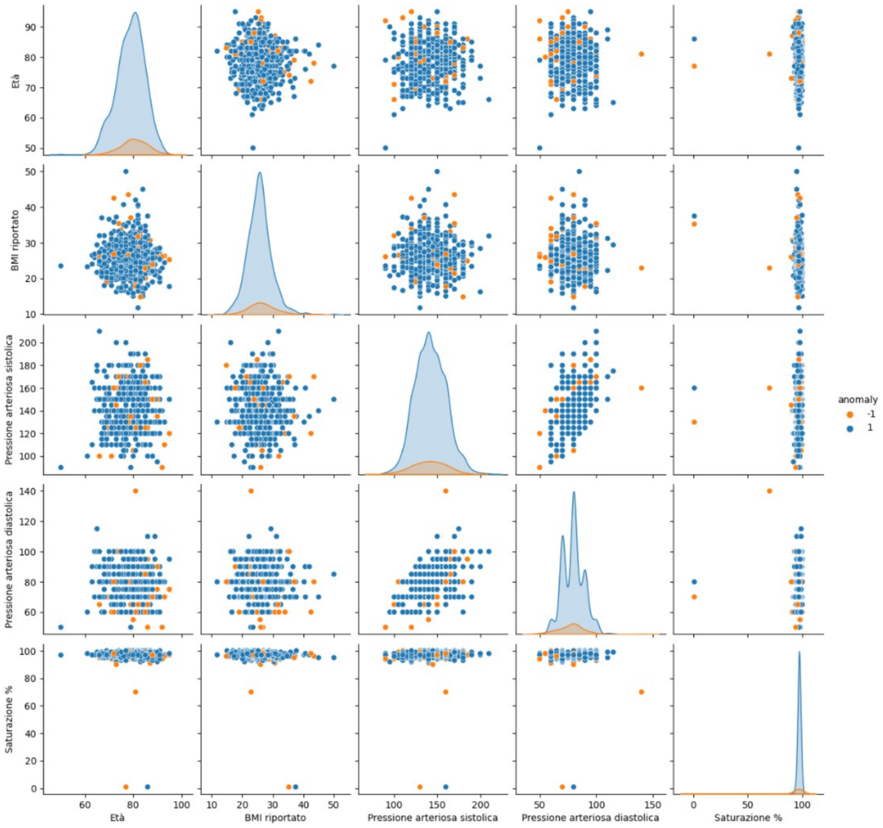

Figure 1: Multivariate outlier graph, that represents the joint distributions of the continuous features for all instances in the dataset: orange dots represent outliers, as identified by the Isolation Forest method, while blue dots represent inliers.

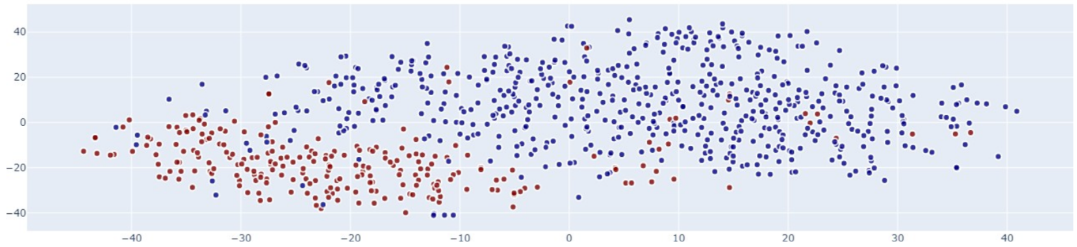

Figure 2: Scatterplot of the dimensionality reduced data. The two axis represent the two axis computed by the t-SNE algorithm. Red dots represent patients associated with mixed dementia, while blue dots represent patients associated with other diagnoses.

To understand why individuals with mixed dementia clustered separately, as highlighted in Figures 2 and 3, we focused on clusters 1 and 2 generated by the HDBSCAN, which were the most populated and representative subsets. We detected significant differences among

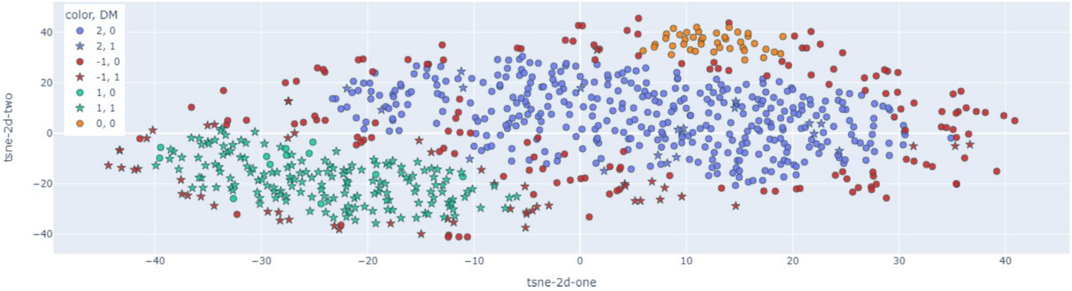

Figure 3: Results of the clustering analysis, as output by the HDBSCAN algorithm. Red dots represent HDBSCAN's noise points (i.e., instances that were in low-density regions of the feature space), while other colors represent clusters discovered by the clustering algorithm.

the two clusters, both for discrete features (Sex: .013, Hypovisus $< .001$, Hearing loss: .020, Atherosclerosis: 0.026, Anxiety/depression: .019, Cerebrovascular disease: <.001, Cognitive impairment: $< .001$, Sleep disorders: .038, Abnormal gait: $< .001$, Peripheral edema: $< .004$) as well as continuous ones (Age: $< .001$, MMSE: $< .001$, Hemoglobin: .009, Red blood cells: .043, Platelets: .024, K: .003, AST GOT: .013, ALT GPT: $< .001$, Vitamin D: $< .001$).

4. Conclusion

In this article we described a high-quality dataset, collected with the aim of providing a repository of information that could be used to explore the characteristics of different neurodegenerative diseases, chiefly among them mixed dementia. Through the application of dimensionality reduction and clustering analysis, as well as statistical inference methods that were used to ground the above-mentioned analyses and assess for relevant differences in characteristics among the identified groups, we provided a phenotypic stratification of the population of patients. Our results highlight how patients affected by mixed dementia can be characterized by means of a phenotypic signature (in terms of both comorbility distribution as well as laboratory chemistry characteristics, see our Discussion on hypothesis testing based on the clustering analysis above) that was markedly different from that of other groups of patients.

While these observations are the results of only a preliminary analysis, we believe them to be very promising in showing the applicability of modern data analysis techniques for addressing the need of establishing more objective, rapid and grounded diagnostic criteria for dementia. Indeed, the features we identified as being characteristic of mixed dementia either refer to easily detectable comorbilities (e.g., hypovisus, hearing loss), or to laboratory parameters that are routinely collected when requesting blood exams [13]. Interestingly, while some of these characteristics have been previously associated with other forms of dementia or neurodegenerative diseases (such as hearing loss [14], blood chemistry parameters [15], or also Vitamin D intake [16]) up to our knowledge this work is the first to provide such a phenotyping for mixed dementia: future clinical examination and analysis should be devoted at better exploring the significance and interpretation of these findings.

In light of our promising results, we believe that future work should be devoted at exploring the potential of applying Machine learning methodologies to the analysis of mixed dementia as well as related diseases. In this sense, we believe that a relevant next step would be the application of predictive modeling and supervised learning techniques for the diagnosis of mixed dementia, as well as the application of eXplainable AI techniques for providing more data-driven exploration of the characteristics of this highly impactful disease. Furthermore, we believe that applying techniques for stratified and adaptive data analysis, to better study the association between population groups and dementia-related diseases.

References

[1] A. Association, 2021 alzheimer's disease facts and figures, Alzheimer's & Dementia 17 (2021) 327–406.

[2] J. T. O'Brien, A. Thomas, Vascular dementia, The Lancet 386 (2015) 1698–1706.

[3] A. Kapasi, C. DeCarli, J. A. Schneider, Impact of multiple pathologies on the threshold for clinically overt dementia, Acta Neuropathologica 134 (2017) 171–186.

[4] C. R. Jack, D. A. Bennett, K. Blennow, et al., Nia-aa research framework: Toward a biological definition of alzheimer's disease, Alzheimer's & Dementia 14 (2018) 535–562.

[5] C. Piller, Blots on a field?, Science (New York, NY) 377 (2022) 358–363.

[6] A. Atri, S. B. Hendrix, V. Pejović, et al., Cumulative, additive benefits of memantine–donepezil combination over component monotherapies in moderate to severe alzheimer's dementia: A pooled area under the curve analysis, Alzheimer's Research & Therapy 7 (2013) 28.

[7] S. Van Buuren, K. Groothuis-Oudshoorn, mice: Multivariate imputation by chained equations in r, Journal of statistical software 45 (2011) 1–67.

[8] F. T. Liu, K. M. Ting, Z.-H. Zhou, Isolation forest, in: 2008 Eighth IEEE International Conference on Data Mining, IEEE, 2008, pp. 413–422.

[9] J. C. Gower, A general coefficient of similarity and some of its properties, Biometrics (1971) 857–871.

[10] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE., Journal of machine learning research 9 (2008).

[11] L. McInnes, J. Healy, S. Astels, hdbscan: Hierarchical density based clustering., Journal of Open Source Software 2 (2017) 205.

[12] C. R. Beam, C. Kaneshiro, J. Y. Jang, et al., Differences between women and men in incidence rates of dementia and alzheimer's disease, Journal of Alzheimer's disease 64 (2018) 1077–1083.

[13] K. Foy, C. Okpalugo, F. Leonard, Usefulness of routine blood tests in dementia work-up, Psychiatric Bulletin 33 (2009) 481–481.

[14] T. D. Griffiths, M. Lad, S. Kumar, et al., How can hearing loss cause dementia?, Neuron 108 (2020) 401–412.

[15] M. Kasa, T. J. Bierma, F. Waterstraat, Jr, et al., Routine blood chemistry screen: a diagnostic aid for alzheimer's disease, Neuroepidemiology 8 (1989) 254–261.

[16] W. B. Grant, Does vitamin d reduce the risk of dementia?, Journal of Alzheimer's Disease 17 (2009) 151–159.